Appendix F. Statistical Methods

Two aspects of the NASS/CDS statistics presented in this report are discussed in this appendix. These aspects are:

- Assigning crash type and general area of damage, and
- The sampling errors for the weighted average counts over the 1994-1996 period.

The method for producing estimates from the 1994-1996 CDS data is to use national ratio-adjusted weights. These sampling weights are appended to the CDS data on the electronic data file. By summing the sampling weights for cases that have a certain characteristic, an estimate of the national total for that characteristic can be produced.

Assigning Crash Type and General Area of Damage

Two different procedures were used to assign General Area of Damage (GAD1) where it was coded as unknown in the NASS/CDS vehicle exterior (VE) form. The first method uses information from elsewhere in the NASS/CDS data files, and the second imputes unknowns based upon a univariate distribution.

1. Assignment based upon other information in the NASS/CDS data files

The SAS program shown below was used to assign GAD1 from the Event file to vehicles where GAD1 is missing of the VE file. The vehicles (towed cars and light trucks) can be classified into three groups with a different approach for each:

- For the first group, damage area was found by matching ACCSEQ1 on the VE form with ACCSEQ on the Event file.
- For the second group, that had no VE form, damage area was found by linking (using VEHNO on the GV file and VEHNUM / OBJCONT on the Event file) to the only damage for that vehicle that was listed on the Event file.
- For the third group, that had no VE form and could not be linked through VEHNUM / OBJCONT, damage area was found by identifying the collision partner in the delta V event, and using ACCSEQ1 from the partner to link to the Event file (this partner was the only other light vehicle in the crash that had delta V estimated by the missing vehicle algorithm for a collision with the vehicle in question).

The numbers of vehicles in the 3 year sample that were assigned GAD1 using this method are summarized as follows:

Table F-1
Cars and Light Trucks with GAD1 Assigned from Event Form, Unweighted Sample Count

Year	method 1	method 2	method 3	Total
1994	79	120	24	223
1995	219	336	69	624
1996	223	385	67	675
Total	521	841	160	1522

SAS program assigning GAD from Event file:

```
*** run this program for each of the years needed, then
combine final data sets (unkn94, unkn95, unkn96), sort
by year psu caseno vehno, and merge with VE file to
```

```
include newly assigned gad info ** ;

libname nass94 'd:\nass\nass94';
options ls=78 pagesize=500;
proc format;
value anydv
   0='DV unknown'
   1='DV known';

value veform
   0='No VE form'
   1='VE form';

value matchveh
   0='No match'
   1='Match';

data gvve1;
* Select towed light passenger vehicles,
  create a 0-1 variable (ANYDV) for delta V availability,
  create a 0-1 variable (VEFORM) for VE form availability;
merge nass94.gv nass94.ve(in=a);
by psu caseno vehno;
if 1<=bodytype<=49;
keep psu caseno vehno
     dvtotal accseq1 gad1 objcont1
     dvbasis anydv veform;
if dvtotal>=0 then anydv=1;
else anydv=0;
veform=a;


data gad_n1 gad_y1;
* Subset to vehicles with delta V estimated
  by the missing vehicle algorithm;
* Creat files for vehicles with GAD1 unknown (GAD_N1)
  and with GAD1 known and from a vehicle impact (GAD_Y1);
* This latter file will be used as a source of ACCSEQ1
  when there is no VE form for the other vehicle;
set gvve1;
if anydv=1 and dvbasis=3;
if gad1=' ' or gad1='9' then output gad_n1;
else if 1<=objcont1<=30 then output gad_y1;

data ve_y1 (keep=psu caseno vehno dvtotal accseq)
     ve_n1 (keep=psu caseno vehno dvtotal);
* From vehicles with delta V but not GAD1,
  create subfiles for those with (VE_Y1) versus
  without (VE_N1) a VE form;
* Rename ACCSEQ1 for merging with the Event file;
set gad_n1;
if veform=1 then do;
   accseq=accseq1;
    output ve_y1;
end;
else output ve_n1;

data vehveh1;
* Create a file with a record for each vehicle-damage
```

```
  combination listed on the Event file for collisions
  between two light passenger vehicles, and
  rename the variables so they can be merged and used
  with the combined GV-VE data selected above;
set nass94.event;
keep psu caseno vehno accseq gadev vehcont;
if 1<=vehnum<=30 and 1<=class1<=49
   then list1='LV';
if (1<=objcont<=30 and 1<=class2<=49) or objcont=70
   then list2='LV';
if list1='LV' and list2='LV';
vehno=vehnum;
gadev=gadev1;
vehcont=objcont;
output;
if list2='LV' then do;
   vehno=objcont;
   gadev=gadev2;
   vehcont=vehnum;
   output;
end;

proc sort data=vehveh1;
by psu caseno vehno accseq;

data ve_y2;
* For vehicles with an estimated delta V, with a VE form,
  with ACCSEQ1 coded, but with GAD1 unknown, merge with
  the Vehicle-Event file using the renamed ACCSEQ variable;
merge ve_y1(in=a) vehveh1;
by psu caseno vehno accseq;
if a=1;

data ve_n2;
* For vehicles with an estimated delta V, but without a VE form,
  merge with the Vehicle-Event file using the renamed VEHNO
variable;
* This produces more than one record per vehicle
  for those vehicles with more than one vehicle-vehicle
collision;
merge ve_n1(in=a) vehveh1;
by psu caseno vehno;
if a=1;

data ve_n_s1(keep=psu caseno vehno
             dvtotal accseq gadev vehcont)
     ve_n_m1(keep=psu caseno vehno
             dvtotal);
 * For vehicles with an estimated delta V, but without a VE form,
  separate into those with only one versus more than one
  vehicle-vehicle collision;
set ve_n2;
by psu caseno vehno;
if first.vehno*last.vehno=1 then output ve_n_s1;
```

```
else output ve_n_m1;

data ve_n_m2;
* Remove redundant entries for the identification
  file of vehicles without a VE form,
  and with more than one vehicle-vehicle collision;
set ve_n_m1;
by psu caseno vehno;
if first.vehno;
matchno=vehno;


data gad_y2;
* Vehicles with delta V estimated by the missing vehicle
algorithm
  and with a VE form are a source of ACCSEQ1 for the other
vehicle
  when that other vehicle has no VE form;
set gad_y1;
keep psu caseno partner matchno accseq;
partner=vehno;
matchno=objcont1;
accseq=accseq1;

proc sort data=gad_y2;
by psu caseno matchno;

data ve_n_m3;
* For vehicles with delta V estimated by the missing vehicle
algorithm,
  without a VE form, and with more than one vehicle-vehicle
collision,
  add ACCSEQ1 from the impacting vehicle involved;
merge ve_n_m2(in=a) gad_y2(in=b);
by psu caseno matchno;
if a=1;
matchveh=b;
drop matchno;
data ve_n_m4 leftovr1;
* For vehicles with an estimated delta V, but without a VE form,
  and with more than one vehicle-vehicle collision,
  and with ACCSEQ1 added from the impacting vehicle,
  separate leftover vehicles with more than one collision
  identified in this manner;
set ve_n_m3;
by psu caseno vehno;
if matchveh=1 and first.vehno*last.vehno=1
    then output ve_n_m4;
else output leftovr1;
data ve_n_m5;
merge ve_n_m4(in=a) vehveh1;
by psu caseno vehno accseq;
if a=1;

data unkn94; set ve_y2 ve_n_s1 ve_n_m5;  year=94;
proc freq data=unkn94; tables gadev /missing;
run;
```

2. **Assignment of crash type and GAD1 based upon imputation**

The unknowns remaining after performing the above procedure were imputed based upon a univariate distribution of crash type. The table below illustrates the steps taken.

Table F-2
Imputation of Unknown GAD1 and Crash Type,
Towed Light Trucks, Total Sample for 1994-1996

row	Crash Type	step 1	step 2	percent
1	rollover 1-3	578	613	11.8%
2	rollover 4+	418	444	8.6%
3	rollover e/e	15	16	0.3%
4	rollover ?	62		
5	**rollover total**	1073	1073	20.7%
6	single front	450	679	13.1%
7	single side	93	140	2.7%
8	single rear	9	14	0.3%
9	single t/u	0	0	0.0%
10	single gad?	281		
11	**single total**	833	833	16.1%
12	multi front	1709	2380	46.0%
13	multi side	497	692	13.4%
14	multi rear	141	196	3.8%
15	multi t/u	2	3	0.1%
16	multi gad?	922		
17	**multi total**	3271	3271	63.2%
18	**total**	5177	5177	100%

The column "step 1" shows the univariate distribution, including unknowns, from the CDS data set after running the SAS program described above. The column "step 2" has the 3 types of unknowns (rollover, type unknown, single vehicle crash with unknown GAD, and multi-vehicle crash with unknown GAD) redistributed into the major crash types according to the proportion of known crash types for that group. For example, the 62 light trucks shown in "step 1", row 4, are redistributed among rows 1 through 3 in “step 2.”

Standard Errors of the CDS Estimates

The national estimates produced from the CDS data may differ from the true values, because they are based on a probability sample of towed cars and not a census of all crashes. The size of these differences may vary depending on which sample was selected. The standard error of an estimate is a measure of the precision or reliability with which an estimate from this particular CDS sample approximates the result of a census.

It is impractical to compute and provide a standard error for each estimate in this report. Instead, generalized standard errors for estimates of totals are presented in the following two tables for vehicle characteristics (Table F3) and for occupant characteristics (Table F4). The generalized standard error tables were produced separately for the vehicle and occupant tables, using three steps.

1. The standard errors for selected estimates in the report were calculated using a Taylor series approximation.

2. An equation that best fit the standard errors was found using regression techniques.

3. Approximate standard errors were generated from this equation, and the generalized standard error tables were produced.

Shown in each table are the values for the estimates and an estimate of one standard error for that value derived from the 1994-1996 CDS data. By adding and subtracting one standard error to the associated estimate, approximate 68% confidence intervals for an estimate can be created. The estimated annual average number of small, crash-involved, towed cars is given in Table 1 of the report as 846,069 cars. To calculate one standard error for this estimate, use Table F3 in this appendix. Since 846,069 does not appear in Table F3, use linear interpolation from the standard error values for the estimates 800,000 and 900,000. One approximate standard error would be 83,400+4,508=87,908. The confidence interval for this estimate would be 846,069 +or- 87,908 or 758,161 to 933,977.

The formula used to compute the standard errors is presented below each table. More information on standard error estimates can be obtained from the National Center for Statistics and Analysis.

Table F3
Crash-Involved Vehicle Characteristics Estimate and Standard Errors

Estimate	Standard Error	Estimate	Standard Error
500	500	100,000	13,600
1,000	700	200,000	24,100
5,000	1,700	300,000	34,200
10,000	2,600	400,000	44,100
20,000	4,100	500,000	53,900
30,000	5,500	600,000	63,800
40,000	6,800	700,000	73,600
50,000	8,000	800,000	83,400
60,000	9,100	900,000	93,200
70,000	10,300	1,000,000	103,100
80,000	11,400	1,100,000	113,000
90,000	12,500	1,200,000	122,900

$$SE = e^{a+b(\ln(x))^2}$$

Where

a= 4.922079
b= 0.034691
x= estimate
SE= Standard Error

Table F4
Crash-Involved Occupant Characteristics Estimate and Standard Errors

Estimate	Standard Error	Estimate	Standard Error
500	300	100,000	19,000
1,000	400	200,000	40,100
5,000	1,200	300,000	63,300
10,000	2,200	400,000	88,300
20,000	4,000	500,000	115,000
30,000	5,800	600,000	143,000
40,000	7,600	700,000	172,500
50,000	9,400	800,000	203,200
60,000	11,300	900,000	235,200
70,000	13,200	1,000,000	268,200
80,000	15,100	1,100,000	302,400
90,000	17,000	1,200,000	337,600

$$\mathrm{SE} = e^{a+b(\ln(x))^2}$$

Where

a= 3.837986
b= 0.04538
x= estimate
SE= Standard Error